FAX

$WE/15033? uf 00

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s)	de Groot, et al.	Examiner:	Graser, J.
Serial No.:	10/049,473	Group Art Unit:	1645
Filed:	July 30, 2002	Docket:	294-120 PCT/US
For:	PNEUMOCOCCAL VACCINES	Dated:	January 7, 2005

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

I hereby certify this correspondence is being deposited with the United States Postal Service as first class mail, postpaid in an envelope, addressed to:
Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313 on January 7, 2005
Signature:_____

DECLARATION UNDER 37 C.F.R §1.132

Sir;

I, Peter Wilhelmus Maria Hermans, declare as follows:

1. I am one of the inventors for the above-referenced patent application.

2. I am associate professor in Molecular Microbiology within the department of Pediatrics of Erasmus Medical Center Rotterdam - Sophia Children's Hospital, The Netherlands. My curriculum vitae is attached as exhibit 1.

3. I am senior scientist (PhD) and head of the laboratory of Pediatrics within the department of Pediatrics of Erasmus Medical Center Rotterdam - Sophia Children's Hospital, The Netherlands.

4. Currently, I am senior staff member of the department of Pediatrics of Erasmus Medical Center Rotterdam - Sophia Children's Hospital, The Netherlands.

BEST AVAILABLE COPY

5.   This invention relates to an isolated protease maturation protein of *S. Pneumoniae*. The protein contains an amino acid sequence as set forth in SEQ. ID NO: 2, and/or a homologous protein thereof.

6.   The term homologous is clearly defined in the specification. Proteins with an E-value (Expect value) of more than 10^{-10}, as determined by Blast or Blastp computer programs, are not considered to be homologous. See the paragraph bridging pages 4 and 5.

7.   According to the National Center for Biotechnology Information (NCBI), accessible thru the internet at the url http://www.ncbi.nlm.nih.gov, the Expect value (E) is defined as:

> ... a parameter that describes the number of hits one can
> 'expect' to see just by chance when searching a database of
> a particular size. It decreases exponentially with the Score
> (S) that is assigned to a match between two sequences.
> Essentially, the E value describes the random background
> noise that exists for matches between sequences. For
> example, an E value of 1 assigned to a hit can be
> interpreted as meaning that in a database of the current size
> one might expect to see 1 match with a similar score simply
> by chance. This means that the lower the E-value, or the
> closer it is to "0" the more "significant" the match is. ...

8.   A copy of the NCBI Blast Frequently Asked Questions (FAQ) which includes the definition of an Expect value is attached as exhibit 2.

9.   Kunsch et al. (WO 98/18930) was cited by the examiner in the Office Action. The examiner alleges that Table 1 of Kunsch et al. discloses a polypeptide having 213 identical amino acids to the claimed SEQ. ID NO: 2. Claimed SEQ. ID NO: 2 is 322 amino acids in length. Thus, the examiner concludes that the polypeptide

of Kunsch et al. with this large number of identical amino acids would inherently be homologous to SEQ. ID NO: 2.
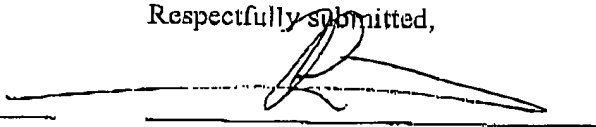
10. Black et al. (U.S. Patent No. 6,348,328 B1) was also cited by the examiner. The examiner alleges that Black et al. teaches a polypeptide which has 48 identical amino acids to the claimed SEQ. ID. NO: 2. The examiner asserts that the fragment of Black et al. containing 48 identical amino acids is a homologous sequence.

11. In the previous Office Action (dated October 29, 2003), a sequence comparison between the Kunsch et al. polypeptide and SEQ. ID NO: 2, and a sequence comparison between the Black et al. polypeptide and SEQ. ID. NO: 2 were included. Exhibit 3 is a copy of the Kunsch et al. sequence comparison. Exhibit 4 contains a copy of the Black et al sequence comparison included in the October 23, 2003 Office Action.

12. The sequence comparison demonstrated a 57.7% match between the amino acid sequence of the Kunsch et al. polypeptide and the amino acid sequence of SEQ. ID NO: 2. See exhibit 3. The sequence comparison showed a 20.3% match between the amino acid sequence of the Black et al. polypeptide and the amino acid sequence of SEQ. ID NO: 2. See exhibit 4.

13. It is well known to those skilled in the art that the computer program used for the sequence comparison is not able to calculate an Expect value for comparisons with non-equal sequence lengths. Therefore, Expect values can only be obtained for sequences with equal lengths.

14. Accordingly, for a polypeptide to be considered homologous to SEQ. ID. NO: 2 in accordance with the specification, the polypeptide must also be the same length as SEQ. ID. NO: 2 since Expect values can only be obtained for sequences with equal lengths.

15.   Further, it would be apparent to one skilled in the art that, even if the proteins being compared were of equal lengths, such a low percentage match (e.g., 57.7% match for the Kunsch et al. polypeptide and 20.3% match for the Black et al. peptide) would not yield an Expect value that is equal to or less than 10^{-10}, as is required in the claimed invention. Therefore, the polypeptides of Kunsch et al. and Black et al. can not be considered to be homologous to SEQ. ID NO: 2, as is required in the claimed invention.

I hereby declare that all statement made herein of my own knowledge are true and that all statements made on information and belief are believed to be true. Further that these statements were made with the knowledge that willfully false statements and the like so made are punishable be fine or imprisonment or both under Section 1001 of Title 18 of the United States Code, and that such willfully false statements may jeopardize the validity of the application of any patent issued thereon.

Respectfully submitted,

Dated: _____ January 7, 2005 _____

Peter Wilhelmus Maria Hermans

BLAST

BLAST Frequently Asked Questions (FAQ)

- News
- Mailing list
- References
- NCBI Contributors

BLAST Services

- FAQs
- Program selection guide
- Web service interface

BLAST Software

- Databases
- Documentation
- Errata
- Executables
- Source code

Support

- Contact us

Tips and Hints:

- Which **BLAST program** should I use?
- How can I search a **batch** of sequences with BLAST?
- How can I write a program to **submit jobs** to NCBI's BLAST servers?
- How can I limit my BLAST search based on **Organism**?
- How can I limit my search to a **subset** of database sequences?
- Is it possible to search for a **motif or pattern** with BLAST?
- How do I perform a similarity search with a **short peptide/nucleotide** sequence?
- Can I use BLAST to compare **two or more sequences** in a multiple sequence alignment?
- What is the **Expect (E) value**?
- What is **low-complexity** sequence?
- Other **Molecular Biology Resources**

Troubleshooting:

- Why do I get the **"No Significant similarity found"** error message?
- Why does my search **timeout** on the BLAST servers?
- Why do I get the error message "ERROR: BLASTSetUpSearch: **Unable to calculate Karlin-Altschul params**, check query sequence"?
- Why do I get the error message "ERROR: **Blast: No Valid Letters to be indexed?"**
- Why do I see a string of **"X"s (or "N")** in my query sequence that I did not put there?
- I have heard that I will be **penalized** if I send a large number of sequences to the servers?

Tips and hints

Q: Which BLAST program should I use?

You have many choices to make between different BLAST programs and databases. Some of these choices are better for answering some questions then others. We have created a selection chart to help you make the decision of BLAST program for the question you are asking. This is the "BLAST Program Selection Guide".

Q: How can I search a batch of sequences with BLAST?

There are three options for "Batch" BLAST searches:
1) **Web MegaBLAST EST analysis tool:** This program is optimized for aligning nucleotide sequences that differ slightly as a result of sequencing or other similar "errors". MegaBLAST is good for scanning a large number of EST type sequences (about 500 kb in length) against large database in search of the closest matches. You can import a file EST sequences in FASTA format or as a list of GenBank accessions or/GIs and have them compared to the BLAST databases. The default is an easily reviewable Hit Table format, although you can download and save the results in Standard pairwise HTML or any of the other result output options. MegaBLAST is available from the BLAST web page, the standalone BLAST executables, or

via the network BLAST client (see below).

2) **Standalone BLAST executables:**The Standalone BLAST executables are command line programs which run BLAST searches against local downloaded copies of the NCBI BLAST databases. The programs will handle either a single large file with multiple FASTA query sequences, or you can create a script to send multiple files one at a time. The executables are available for a wide variety of platforms, including many "flavors" of UNIX (LINUS, Solaris, etc.) Windows PC and even Mac OSX.

The Standalone executables are available at the anonymous FTP location: ftp://ftp.ncbi.nih.gov/blast/executables/ There is information on the Standalone BLAST executables available in the README file at ftp://ftp.ncbi.nih.gov/blast/documents/blast.txt which is also bundled with the downloaded binaries.

3) **BLAST Network Client 'blastcl3':**The BLAST 2.0 Network client will allow you to submit a single file of FASTA sequences over an internet connection to the NCBI BLAST databases. You submit searches through the client to the NCBI servers and do not need to download the database locally. The BLAST Network client executables are located at: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/ There are blastcl3 executables for various UNIX platforms, PC Windows and Macintosh.

Q: How can I write a program to submit jobs to NCBI's BLAST servers?

By using the URLAPI. Documentation also available in postscript and PDF.

Q: How can I limit my BLAST search based on Organism?

The option to limit a search to organism and even taxonomic classification is part of the "Limit by Entrez Search" option on most standard BLAST search pages. There is a pull down menu to select the most common organisms found in GenBank and also a field to input the species name, or classification (example: "eubacteria"). Using this option will cause your query sequences to be compared only to sequences in our databases from that organism.

There are also several "specialized" BLAST Pages devoted to different organisms on the main BLAST web page.

How can I limit my search to a subset of database sequences?

You can use the "Limit by Entrez Search" option found on most Standard BLASTR search pages to run an Entrez search and have your query sequence compared to the resutls of this search. For example, if you wanted to limit you search to all phosphorylase sequences from mouse you could enter the following valid Entrez search strategy in the Limit by Entrez field of the BLAST search page: phosphorylase AND "Mus musculus"[Organism]

Q: Is it possible to search for a motif or pattern with BLAST?

There are two general approaches to this type of questions. First do you wish to find if motifs exist in your query sequence, or do you have a known motif and wish to find other protiens or nucleotides with this motif?

In the first case, finding motifs in your query sequence can be done for proteins using the CDD (Conserved Domain Database) and CDART (Conserved Domain Architecture Retrieval Tool) tools. CDD allows you to compare your protein to an database of alignments and profiles

representing protein domains conserved in molecular evolution as well as 3-dimensional protein structures in the MMDB database. These tools use popular protien motif databases, PFam (http://pfam.wustl.edu/) and Smart (http://smart.embl-heidelberg.de) in addition to the MMDB database.

For conditions of the second case if you have a known motif and wish to identify other proteins with this motif you can use PHI-BLAST. PHI-BLAST searches take a motif pattern and protein sequence as input and then compares these to the NCBI protein databases looking for other proteins which contain conserved regions similar to the motif entered.

For nucleotides it is only possible to search with short query sequences representing your motif or region of interest with the Nucleotide BLAST "Search for short nearly exact matches" service from the main BLAST web page. This can find other sequences whicvh contain similar nucleotide patterns. however there are no database of nucleotide patterns which can identify patterns in your nucleotide query sequence.

You may also be interested in checking out other molecular biology web sites, such as those mentioned in the Other Molecular Biology Resources section at the end of this FAQ, for motif searching software.

Q: How do I perform a similarity search with a short peptide/nucleotide sequence?

There is a special page with pre-set parameters for searching with short sequences. You can access this page by clicking the "Search for short nearly exact matches" link on the main BLAST web page.

Essentially for these searches, the Expect value has been increased and the word size decreased to optimise for short hits which generally score a large E value require smaller word sizes to initiate formation of the HSP for extension. In addition, for proteins, the matix "PAM30" becomes the default which optimises hits to smaller sequences which have a lower percentage of evolutionary drift in general.

Q: Can I use BLAST to compare to two or more sequences in a multiple sequence alignment?

You can use the BLAST 2 Sequences service to compare two nucleotide or two protein sequences against each other using the Gapped BLAST algorithm. The this will allow you to perform a BLAST search between the two sequences allowing for the introduction of gaps (deletions and insertions) in the resulting alignment. Remember that BLAST is a "local" alignment program and does not make global alignments between sequences to calculate total percent homologies.

To compare one sequence against a specific sequence or set of sequences, you can also use a separate multiple sequence alignment program. There are many such software tools available to do this. You may also be interested in checking out other molecular biology web sites, such as those mentioned in the Other Molecular Biology Resources section at the end of this FAQ.

Q: What is the Expect (E) value?

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. It decreases exponentially with the Score (S) that is assigned to a match between two sequences. Essentially, the E value

describes the random background noise that exists for matches between sequences. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0" the more "significant" the match is. However, keep in mind that searches with short sequences, can be virtually indentical and have relatively high EValue. This is because the calculation of the E-value also takes into account the length of the Query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance. For more details please see the calculations in the BLAST Course.

The Expect value can also be used as a convenient way to create a significance threshold for reporting results. You can change the Expect value threshold on most main BLAST search pages. When the Expect value is increased from the default value of 10, a larger list with more low-scoring hits can be reported.

Q: What is low-complexity sequence?

Regions with low-complexity sequence have an unusual composition andthis can create problems in sequence similarity searching (Wootton & Federhen, 1996). Low-complexity sequence can often be recognized by visual inspection. For example, the protein sequence PPCDPPPPPKDKKKKDDGPP has low complexity and so does the nucleotide sequence AAATAAAAAAAATAAAAAAT. Filters are used to remove low-complexity sequence because it can cause artifactual hits (please also see Q: After running a search why do I see a string of "X"s (or "N"s) in my query sequence that I did not put there?)

In BLAST searches performed without a filter, often certain hits will be reported with high scores only because of the presence of a low-complexity region. Most often, this type of match cannot be thought of as the result of homology shared by the sequences. Rather, it is as if the low-complexity region is "sticky" and is pulling out many sequences that are not truly related.

Other Molecular Biology Resources:

The on-line **BLAST Course** was written by Dr. Stephen Altschul and discusses the basics of the Gapped BLAST algorithm. In addition the **full text** of the 1997 Nucleic Acids Research paper "**Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**" is also available on-line.

Other links:
     European Bioinformatics Institute (EBI) BioCatalog
     Indiana University IUBio Archive
     Sequence manipulation site

Troubleshooting

Q: Causes for "No significant similarity found".

Below are several reasons that a BLAST search can result in the "No significant similarity found" message.

Short Sequences: There is a special BLAST optimized for searchig with small sequences. Go to tbe main BLAST web page and select the "**Search for short nearly exact matches**" link for Nucleotide - Nucleotide or Protein Protein sections.

Filtering: BLAST filters regions of low-complexity (for a description of low-complexity see "What is low-complexity sequence?" below). If your sequence contains large regions of "low complexity" it may not significant hits to the database. You can turn off filtering by setting the "**Filter**" option to "**None**" using the pull down tab.

Query Format: Another reason you may see the "No Significant Similarity found" message is using the wrong type of sequence in your search.

1) Accession/GI Number or FASTA. Check that you have the Input Data set to the correct format for your Query. Set the pull down menu to "Accession number or Gi" to search with GenBank accession numbers or Gi numbers. Set to FASTA for raw amino acid or nucleotide sequences. For more information on FASTA format, click here.

2) Sequence type and Program combination. You can search with an amino acid query sequence using the blastp and tblastn programs. With nucleotide query sequences you can use blastn, blastx, and tblastx. Please note that tblastx program cannot be used with the nr database on the BLAST Web page.

For more information on the BLAST programs, click here.

Q: Why does my search timeout on the BLAST servers?

Certain combinations of BLAST searches with large sequences against large databases can cause the BLAST servers to timeout. This has to do with a limit on the server CPU's which prevents sequences which generate many HSPs from hoarding server resources.

However there are some things you can do to prevent timeout and generate results from large sequences.

- Some sequences contain large regions of ALU repeats. In this case you can select the "Human Repeat" filtering option on the main BLAST search page. This will mask repeat regions which generate a large number of biologically uninteresting hits to the databases.

- Increase the Word Size to 20 - 25. With a default Word Size of 7, the BLAST algorithm finds initial HSPs of 7 bases in length and begins extension of these from either end. In a large sequence this can generate 100's of initial HSPs between the query sequence and even a single large genomic sequence in the databases. Increasing the Word Size to 25 makes the initial HSP smaller, limiting the number small initial fragments to be extended.

- Decrease the Expect value to 1.0 or lower. Many hits from large sequences are to many small fragments in the database. The expect value for these searches is such that decreasing the expect value will eliminate these results, and concentrate on results which are more likely to contain large coding regions and genomic fragments.

If you are still seeing a "timeout" error message after making the above changes, please contact blast-help@ncbi.nlm.nih.gov with the RID of your search.

Q: Why do I get the message "ERROR:BLASTSetUpSearch: Unable to calculate Karlin-Altschul params, check querysequence" ?

This will happen if your entire query sequence has been masked by low complexity filtering. You will need to turn filtering off to get hits. For further information on filtering, please read the sections of the BLAST FAQs on Q: What is low-complexity sequence? and also Q: After

Q: Why do I get the message "ERROR: Blast: No valid letters to be indexed"?

You may have accidentally entered an accession number in the search box without changing the input selection from "Sequence in FASTA format" to "Accession or gi". You will also see this error message if too many ambiguity codes (R,Y,K,W,N, etc. fornucleotides) are present in your query sequence. Although BLAST allows ambiguity codes, be aware that these will always contribute a negative score in nucleic acid searches. Thus, sequences such as degenerate PCR primers with ambiguity codes maynot find any significant hits even though they may be designed from sequences that are present in the database.

Q: After running a search why do I see a string of "X"s (or "N"s) in my query sequence that I did not put there?

You are seeing the result of automatic filtering of your query for low-complexity sequence that is performed to prevent artifactual hits. The filter substitutes any low-complexity sequence that it finds with the letter "N" in nucleotide sequence (e.g., "NNNNNNNNNNNNN") or the letter "X" in protein sequences (e.g., "XXXXXXXXX"). Low-complexity regions can result in high scores that reflect compositional bias rather than significant position-by-position alignment (Wootton amp; Federhen, 1996). Filter programs can eliminate these potentially confounding matches from the blast reports, leaving regions whose BLAST statistics reflect the specificity of their parities alignment. Queries searched with the blastn program are filtered with DUST. The other BLAST programs use SEG.

Q: How can I see low-similarity matches when there are many strong hits to my query sequence? Often, when the query is a member of a large sequence family, the summary hit list and the alignments returned only contain very high scoring hits. To look at low-similarity matches, you must increase the maximum number of results returned. On the BLAST Web pages, often it is sufficient to increase the size of the summary hit list and the number of alignments shown using the menus on the Advanced pages. However, it is possible to increase the lists even further using the Other Advanced Options box on the Advanced BLAST pages. For BLAST 2.0, "-v 2000", for example, will increase the number of descriptions returned in the summary hit list to 2000. The option "-b 2000" will similarly increase the number of alignments returned.

Q: I have heard that I will be penalized if I send a large number of sequences to the servers? .

The NCBI WWW BLAST server is a shared resource and it would be unfair for a few users to monoplize it. To prevent this, the server keeps track of how many queries are in the queue for each user and penalzies those users with many queries in the queue. This is done by calculating a 'Time of Execution' (TOE). If a user has only one query in the queue, then the TOE is set to the current time. As a user adds more queries to the queue, then the TOE is set to the current time, plus 60 seconds for every query in the queue. An example would be if a user sent in five requests one after the other without waiting for any to be worked on, then the TOE's for the requests would be:

1st request: current time
2nd request: current time + 60 seconds
3rd request: current time + 120 seconds

4th request: current time + 180 seconds

5th request: current time + 240 seconds

The BLAST server works through requests in the order of earliest to latest TOE. A query will be executed before it's TOE, if there are no other queries with an earlier TOE. Users with large numbers of queries are encouraged to use the BLAST servers at off-peaks hours, which are from 8 p.m. to 8 a.m. (EST).

```
Qy  151 YTPDVTAQIIRLNNEDKAKEVLEKAKAEGADFAQLAKDNSTDEKTKENGGEITFDSASTE 210
        ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db  121 YTPDVTAQIIRLNNEDKAKEVLEKAKAEGADFAQLAKDNSTDEKTKENGGEITFDSASTE 180

Qy  211 VP-EQVKKAAFA------LDVD 225
        ||  ||||        ||||
Db  181 VPGASPKKPLFAFRCGMVFLDVD 203
```

RESULT 5
ABP54573

```
ID  ABP54573 standard; Protein; 213 AA.
XX
AC  ABP54573;
XX
DT  04-SEP-2002  (first entry)
XX
DE  S. pneumoniae SP021 protein sequence SEQ ID NO:34.
XX
KW  Streptococcus pneumoniae; epitope; vaccine; antigenic protein;
KW  antibacterial; Streptococcal infection; detection.
XX
OS  Streptococcus pneumoniae.
XX
PN  US2002061545-A1.
XX
PD  23-MAY-2002.
XX
PF  22-JAN-2001; 2001US-0765272.
XX
PR  30-OCT-1997;  97US-0961083.
XX
PA  (CHOI/) CHOI G H.A.
PA  (KUNS/) KUNSCH C A.
PA  (BARA/) BARASH S C.
PA  (DILL/) DILLON P J.
PA  (DOUG/) DOUGHERTY B.
PA  (FANN/) FANNON N R.
PA  (ROSE/) ROSEN C A.
XX
PI  Choi GH, Kunsch CA, Barash SC, Dillon PJ, Dougherty B, Fannon MR;
PI  Rosen CA;
XX
DR  WPI; 2002-479261/51.
DR  N-PSDB; ABQ84808.
XX
PT  New Streptococcus pneumoniae antigens, useful for detecting
PT  Streptococcus and for preventing or attenuating disease caused by
PT  Streptococcus infection -
XX
PS  Claim 11; Page 24; 70pp; English.
XX
CC  ABQ84792 to ABQ84904 represents nucleic acids which encode the
CC  Streptococcus pneumoniae antigens given in ABP54557 to ABP54669.
CC  The S. pneumoniae antigens have antibacterial activity and can be
CC  used in vaccines. The S. pneumoniae antigens can also be used to
CC  prevent or attenuate a Streptococcal infection in an animal. The
CC  polynucleotides encoding the S. pneumoniae antigens can be used to
CC  detect Streptococcus nucleic acids. ABQ84905 to ABQ85130 represent
CC  Streptococcus pneumoniae ORFs (open reading frames)
CC  primers used in the cloning of S. pneumoniae ORFs (open reading frames)
CC  which are used in an example from the present invention.
```

Sequence 213 AA;

```
           57.7%    Score 916; DB 23; Length 213;
Similarity 93.6%;   Pred. No. 3.3e-62;
Conservative  0;    Mismatches  5;   Indels  8;   Gaps  2;
```

```
b  241 KKTEKSSNIDDYKEILKTVILTQKQNDSTFVQSIIGKELQAANIKVKDQAFQNIFTQYIG 300
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
y  310 GGDSSSSSTSNE 322
       ||||||||||||
b  301 GGDSSSSSTSNE 313
```

RESULT 4
AW55079

```
D  AAW55079 standard; Protein; 213 AA.
XX
C  AAW55079;
XX
T  OCT-1998  (first entry)
XX
E  Streptococcus pneumoniae SP0021 protein.
XX
W  Streptococcus pneumoniae; antigen; vaccine; infection; diagnosis;
W  detection; pneumonia; otitis media; meningitis.
XX
S  Streptococcus pneumoniae.
XX
N  WO9818930-A2.
XX
D  07-MAY-1998.
XX
F  30-OCT-1997;  97WO-US19422.
R  31-OCT-1996;  96US-0029960.
XX
A  (HUMA-) HUMAN GENOME SCI INC.
XX
I  Choi GH, Hromockyj A, Johnson LS, Kunsch CA;
XX
R  WPI; 1998-272224/24.
R  N-PSDB; AAV27340.
XX
T  Nucleic acid encoding antigenic peptide(s) from Streptococcus
T  pneumoniae - or their epitope-containing fragments, useful in
T  protective or therapeutic vaccines, and for diagnosis
XX
S  Claim 11; Page 55; 118pp; English.
XX
C  The present sequence represents a protein from Streptococcus pneumoniae.
C  The nucleic acid sequence encoding the Streptococcus pneumoniae protein
C  can be useful in vaccines for inducing protective antibodies against
C  Streptococcus pneumoniae, for treatment or prevention of infection e.g.
C  pneumonia, otitis media or meningitis. Probes based on the nucleic acid
C  are used to detect Streptococcus infection (by usual hybridisation or
C  amplification methods), also for isolating Streptococcus genes or their
C  allelic variants. The protein can be used similarly to detect specific
C  antibodies in standard immunoassays, especially for diagnosing or
C  monitoring infections. Antibodies which bind the protein are used to
C  detect corresponding antigens, to purify the protein and for passive
C  immunisation (optionally coupled to a toxin). Vaccines are administered,
C  e.g. by injection, orally or through the skin, typically at 0.01-1000
C  (especially 10-300) mu g/ml per dose.
XX
SQ  Sequence  213 AA;
```

```
Query Match          57.7%;   Score 916; DB 19; Length 213;
Best Local Similarity 93.6%;  Pred. No. 3.3e-62;
Matches 190;  Conservative  0;  Mismatches  5;  Indels  8;  Gaps  2;
```

```
Qy  31 SKGSEGADLISMKGDVITEHQFYEQVKSNPSAQQVLLNMTIQKVFEKQYGSELDDKEVDD 90
       |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db   1 SKGSEGADLISMKGDVITEHQFYEQVKSNPSAQQVLLNMTIQKVFEKQYGSELDDKEVDD 60

Qy  91 TIAEEKKQYGENYQRVLSQAGMTLETRKAQIRTSKLVELAVKKVAEAELTDEAYKKAFDE 150
       |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db  61 TIAEEKKQYGENYQRVLSQAGMTLETRKAQIRTSKLVELAVKKVAEAELTDEAYKKAFDE 120
```

Qy   70 TIQKVFEEKQYGSELDDKEVDDTIAEEKKQYGENYQRVLSQAGMTLETRRKAQIRTSKLVEL 129
              : | : || || : ||| : |:: ||: | : || : | || | ||
Db   70 IVFKAFDDVYGKEISSKEVTKEYNEQIKKLGPNYKEQLKGPNYKEQLKAKEVLEKAKAEGADFAQLAKDN 189

Qy   130 AVKKVAEAELTDEAYKKAFDEYTPDVTAQIIRLNNEDKAKEVLEKAKAEGADFAQLAKDN 189
              :| : ||| : | :||: || | : |: ||
Db   130 GLK--ANVKLTDKDLDTAWKEFYPEVSTQIILFSTKEEADKAKKEAN-EGENFSKLVQAY 186

Qy   190 STDEKTKENGGEITFDSASTEVPEQVKKAAFALDVDGVSDVITATGTQAYSSQYYIVKLT 249
              : | :||| || : | :||| | ||:| ||: | : |||: : : ||:
Db   187 GKNKTLKETDGKMNFDSTNPEIPTEVKAAFKLKNGEVSDIIPVTDPTTYQQSYLVKMV 246

Qy   250 KKTEKSSNIDDYKEKLKTVILTQKGNDSTFVQSIIGKELQAANIKVKDQAFQNIF 304
              || : || || : || : || :|: : || | :| ||
Db   247 KKQDKGSNKDKYKSELEKIATEAKTLDTEFMDKTIRKVMKKDNVTIKDPYVKNIF 301

RESULT 8
US-08-858-207A-508
; Sequence 508, Application US/08858207A
; Patent No. 6348328
; GENERAL INFORMATION:
; APPLICANT: Black, Michael
; APPLICANT: Hodgson, John
; APPLICANT: Knowles, David
; APPLICANT: Nicholas, Richard
; APPLICANT: Stodola, Robert
; TITLE OF INVENTION: No. 6348328el Compounds
; NUMBER OF SEQUENCES: 552
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: SmithKline Beecham Corporation
; STREET: 709 Swedeland Road
; CITY: King of Prussia
; STATE: PA
; COUNTRY: USA
; ZIP: 19406-0939
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Diskette
; COMPUTER: IBM Compatible
; OPERATING SYSTEM: DOS
; SOFTWARE: FastSEQ for Windows Version 2.0
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/858,207A
; FILING DATE: 09-MAY-1997
; CLASSIFICATION: 435
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: 60/017670
; FILING DATE: 14-MAY-1996
; ATTORNEY/AGENT INFORMATION:
; NAME: Gimmi, Edward R
; REGISTRATION NUMBER: 38,891
; REFERENCE/DOCKET NUMBER: P50475
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 610-270-4478
; TELEFAX: 610-270-5090
; TELEX:
; INFORMATION FOR SEQ ID NO: 508:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 74 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: No. 6348328e
US-08-858-207A-508

Query Match   20.3%;   Score 322; DB 4; Length 74;
Best Local Similarity   97.1%;   Pred. No. 1.2e-19;
Matches   66; Conservative   1; Mismatches   1; Indels   0; Gaps   0;

Qy   10 MKKKLLAGAITLLSVATLAACSKGSEGADLISMKGDVITEHQFYEQVKSNPSAQQVLLNM 69
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db   1 MKKKLLAGAITLLSVATLAACSKGSEGADLISMKGDVITEHQFYEQVKNNFSAQXVLLNM 60

Qy   70 TIQKVFEK 77
              ||||||||
Db   61 TIQKVFEK 68

RESULT 9
US-09-134-001C-3811
; Sequence 3811, Application US/0913400IC
; Patent NO. 6380370
; GENERAL INFORMATION:
; APPLICANT: Lynn Doucette-Stamm et al
; TITLE OF INVENTION: NUCLEIC ACID AND AMINO ACID SEQUENCES RELATING TO STAPHYLOCOCC
; TITLE OF INVENTION: EPIDERMIDIS FOR DIAGNOSTICS AND THERAPEUTICS
; FILE REFERENCE: GTC-007
; CURRENT APPLICATION NUMBER: US/09/134,001C
; CURRENT FILING DATE: 1998-08-13
; PRIOR APPLICATION NUMBER: US 60/064,964
; PRIOR FILING DATE: 1997-11-08
; PRIOR APPLICATION NUMBER: US 60/055,779
; PRIOR FILING DATE: 1997-08-14
; NUMBER OF SEQ ID NOS: 5674
; SEQ ID NO 3811
; LENGTH: 330
; TYPE: PRT
; ORGANISM: Staphylococcus epidermidis
US-09-134-001C-3811

Query Match   13.2%;   Score 210; DB 4; Length 330;
Best Local Similarity   27.5%;   Pred. No. 1.4e-09;
Matches   95; Conservative   55; Mismatches   112; Indels   84; Gaps   20;

Qy   23 SVATLAAC-SKGSEGAD--LISMK-GDV----ITEHQFYEQVKSNPSAQQVLLNMTIQKV 74
              || | ||| | || : || | |:|| | || : :| : ||
Db   19 SALLLGACGSNATESKDNTLISSKAGDVKVADVMKKMGKEQI-ANTS----FSIVLNKV 72

Qy   75 FEKQYGSELDDKEVDDTIAEEKKQYG--ENYQRVLSQAGMTLETRKAQIRTS---KLVEL 129
              : : ||:||| | :| :: ||||| : :| | |:|| | : : :| |
Db   73 LADKYKDKVDTKDIDKDIKKEEKQYGGKDQFESMLKQQGMSLDDYKEQKKLSAYQKQLLL 132

Qy   130 AVKKVAEAELTDEAYKKAFDEYTPDVTAQI--IRLNNED------KAKEVLEKAKAE- 178
              || : || | | : |||| || ||||| |||| | :
Db   133 DKVNVSDKEIKENSKK-----------TSHILIVKSKSSDKEGLSDKKAKEAKEIQKEV 182

Qy   179 --GADFAQLAKDNSTDEKTKENGGEITFDSASTEVPEQVKKAAFALDVDGVSDVI-TAT 234
              | | | |:| : : | : : | ||: | | | | | | ||
Db   183 EKNPNKFGEJAKKESMDSSSAKKDGSLGY-VIKGQMVDSFEKALFKLKEGEVSKVVKTDY 241

Qy   235 GTQAVSSQYYIVKLTKKTE--KSSNI-----------DDYKEKLKTVILTQKQ 274
              | | | | :| | :| ||:| | : || | | | :
Db   242 G------YHIIKADKETDFNSEKSNIKQKLIEEKVQKPKLLTDAYKELLKEYKVDYKD 294

Qy   275 ND--STFVQSIIGKELQAANIKVKDQAFQNIFTQYIGGDSSSSSS 318
              || | : || || ::| :   ::| :| | ||
Db   295 RDIKKMEDSILDPD------KIKQQQQ----QSQGGSGLTNSGS 330

RESULT 10
US-09-252-991A-17522
; Sequence 17422, Application US/09252991A
; Patent No. 655,795
; GENERAL INFORMATION:
; APPLICANT: Marc J. Rubenfield et al.
; TITLE OF INVENTION: NUCLEIC ACID AND AMINO ACID SEQUENCES RELATING TO PSEUDOMONAS
; TITLE OF INVENTION: AERUGINOSA FOR DIAGNOSTICS AND THERAPEUTICS
; FILE REFERENCE: 107196.136
; CURRENT APPLICATION NUMBER: US/09/252,991A
; CURRENT FILING DATE: 1999-02-18
; PRIOR APPLICATION NUMBER: US 60/074,788
; PRIOR FILING DATE: 1998-02-18
; PRIOR APPLICATION NUMBER: US 60/094,190
; PRIOR FILING DATE: 1998-07-27
; NUMBER OF SEQ ID NOS: 33142
; SEQ ID NO 17422